Labour Economics 26 (2014) 1–10

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

The co-twin methodology and returns to schooling — testing a critical assumption ☆☆

Örjan Sandewall [a,*], David Cesarini [b,c,d], Magnus Johannesson [e]

[a] AlixPartners LLP, 20 North Audley Street, London W1K 6WE, United Kingdom
[b] Center for Experimental Social Science, New York University, 19 W. 4th Street, 6FL, New York, NY 10012, United States
[c] Division of Social Science, New York University Abu Dhabi, United Arab Emirates
[d] Research Institute of Industrial Economics, Stockholm, Sweden
[e] Department of Economics, Stockholm School of Economics, PO Box 6501, Sveavägen 65, SE-113 83, Sweden

HIGHLIGHTS

- An influential literature on returns to schooling uses data on identical twins.
- These studies assume that twins are identical as to relevant underlying abilities.
- Using a detailed novel dataset, we find strong evidence against this assumption.
- Adolescent IQ differences significantly add to within-pair schooling-wage equations.
- IQ differences reduce the within-pair estimated returns to schooling by about 15%.

ARTICLE INFO

Article history:
Received 30 July 2013
Received in revised form 12 October 2013
Accepted 30 October 2013
Available online 7 November 2013

JEL Classification:
I21
J24

Keywords:
Returns to schooling
Twins
Equal ability assumption

ABSTRACT

Twins-based estimates of the return to schooling have featured prominently in the economics of education literature. Their unbiasedness hinges critically on the assumption that within-pair variation in schooling is explained by factors unrelated to wage earning ability. This paper develops a framework for testing this assumption and shows, in a large sample of monozygotic twins, that the twins-based estimated return to schooling falls if adolescent IQ test scores are included in the wage equation. Using birth weight as an alternative proxy for ability yields qualitatively similar results. Our results thus cast doubt on the validity of twins-based estimates.

1. Introduction

Knowledge about the causal effect of schooling on earnings and other economic outcomes has important implications for educational policy, for efforts to better understand the evolution of inequality and for studies examining the sources of economic growth (Card, 2001; Katz and Autor, 1999). Yet, it has long been known that efforts to obtain precise estimates of the causal effect of schooling on earnings are complicated by the endogeneity of schooling decisions. In particular, there is a widely shared view that estimates of the marginal returns to schooling will be biased unless proper account is taken of heterogeneities in latent ability. If the propensity to invest in further years of education is also

☆☆ We thank the editor (Ian Walker), two referees, John van Reenen, Fabian Waldinger and the seminar audience at London School of Economics for their helpful comments. We are especially grateful to Steve Pischke for providing constructive feedback on several earlier drafts of the paper. This research was funded by the Hedelius Wallander Foundation, the Swedish Council for Working Life and Social Research (FAS), and the Swedish Research Council. We take full responsibility for any remaining errors. The opinions expressed are those of the author(s) and do not necessarily reflect the views of AlixPartners, LLP, its affiliates, or any of its or their respective other professionals or clients.

* Corresponding author. Tel.: +44 7825 625183.
E-mail addresses: osandewall@alixpartners.com (Ö. Sandewall), dac12@nyu.edu (D. Cesarini), Magnus.Johannesson@hhs.se (M. Johannesson).

0927-5371
http://dx.doi.org/10.1016/j.labeco.2013.10.002

directly related in a positive way to the ability to earn wages, then this will cause an upward bias in estimates of the effect of an additional year of schooling on wages (see for example Card, 1999, 2001).

A number of approaches to eliminating or mitigating this endogeneity problem have been proposed. One strand of work uses instrumental variable analysis to try to reduce the bias of the estimates (Angrist and Krueger, 1991; Card, 2001).[1] A second influential strand of the literature has exploited within-family variation in general, and variation within monozygotic (MZ) twin pairs in particular, to try to control for unobserved ability. Under the key identifying assumption that within-family variation in schooling is explained by factors unrelated to wage earning ability, resulting estimates are consistent as long as problems of measurement errors in the schooling variable can be dealt with adequately. If two siblings have identical abilities, then the identifying assumption is of course satisfied. Especially with regards to MZ twins, the attraction of the assumption of equal ability is easily understood. MZ twins are the result of a fertilized egg splitting in two shortly after conception, resulting in two identical individuals who are virtually identical genetically (Martin et al., 1997). Furthermore, MZ twins (or "identical" twins, as they are often referred to) are typically raised by the same parents, go to the same school, and are influenced by the same peer groups when growing up.

In labor economics, twins-based estimates of the return to schooling have featured prominently; see, for instance, the survey in Card (1999). A string of papers applying co-twin methodology have been published in prominent economic journals (Ashenfelter and Krueger, 1994; Behrman, Rosenzweig, and Taubman, 1994; Ashenfelter and Rouse, 1998; Miller, Mulvey, and Martin, 1995; Bonjour et al., 2003; Amin, 2011) as well as field journals (Isacsson, 1999; Behrman and Rosenzweig, 1999; Rouse, 1999; Isacsson, 2004; Miller et al., 2006; Zhang, Liu, and Yung, 2007).[2]

The idea that the latent wage earning ability of two individuals in a pair of identical twins would be virtually identical is not hard to accept, a priori. However, identical ability begs the question of what causes observed within-pair differences in schooling, as standard optimizing models predict that two identically able individuals would choose the same level of schooling (Ashenfelter and Rouse, 1998; Becker, 1964; Ben-Porath, 1967). Any observed variation in schooling must then be explained by "optimizing errors", or differences in preferences for schooling which do not affect wage earning ability. Hence, it is assumed that differences in schooling across the population are caused by ability differences, but that this is not true within twin pairs.

A natural hypothesis is that within-pair variation in ability may explain within-pair variation in schooling, thereby violating the assumption of "optimization error".[3] This potential problem with the co-twin methodology was first demonstrated by Griliches (1979); although twins may have very similar levels of ability, the observed similarities in years of schooling and income are also large. Therefore, even though within-pair differences are purged from most of the heterogeneities in ability, they may also lack most of the useful variation in schooling and income. Griliches (1979) noted that when the degree of twin similarity is the same for ability and for schooling, first-differencing contributes nothing in terms of removing ability bias. This critique has been further developed both conceptually and empirically by Neumark (1999) and Bound and Solon (1999), who also point out that a priori the relationship between the degrees of similarity in ability and schooling, respectively, is not clear.

The contribution of this paper is to provide results from empirical assessments which rely on less restrictive assumptions than previous tests in the literature, and which use better proxies for ability than has generally been employed. To this end, we use a large sample of Swedish male pairs of MZ twins. Our data contain information on income, adolescent IQ, birth weight, and two separate measures of schooling. The dataset is distinguished from previous studies as it includes dual measures of schooling as well as appropriate ability measures and that we directly examine how controlling for proxies for ability in a standard co-twin wage regression affects the estimated return to schooling. The main findings of the paper are that (i) within-pair differences in IQ test scores are significantly associated with income even after accounting for differences in schooling, (ii) that within-pair differences in IQ test scores have a statistically and economically significant effect on within-pair differences in schooling, and (iii) controlling for IQ test scores reduces within-pair estimates of returns to schooling by about 15% across various specifications and variable definitions.

These results cast doubts on the validity of the co-twin approach to estimating the returns to schooling, and provide some additional empirical evidence for the critique of within-family estimation advanced by Griliches (1979), Bound and Solon (1999), Neumark (1999) and others. The evidence reported here suggests that the quasi-experiment of MZ twinning does not approximate the ideal experiment, namely random assignment of educational attainment holding ability and other background factors constant, particularly well. In fact, under plausible assumptions about the reliability ratio of the within-pair difference in adolescent IQ and educational attainment, the within-pair correlation between IQ and schooling is about 0.30.

Our results are also complementary to a recent economics literature (e.g. Behrman and Rosenzweig, 2004; Black et al., 2007; Royer, 2009) which documents convincingly in large samples that the within twin pair difference in birth weight – a commonly used proxy for the quality of the prenatal environment – predicts outcomes such as intelligence, earnings and educational attainment. These papers, whilst not framed directly as an attempt to interrogate the "equal ability assumption", provide strong suggestive evidence that the key identifying assumption in twins-based estimates of the return to schooling is violated.[4] They do not allow us to determine the extent to which birth weight acts on income directly, rather than through schooling, and hence leave open the question of whether it is the non-ability or the ability components of schooling which differ between twins.

An additional concern about twins-based estimates relates to measurement error in schooling. As was noted by one of the first authors to apply this methodology (Taubman, 1976), differencing within pairs will usually decrease the signal to noise ratio, and hence serves to exacerbate the problem of imperfectly observed schooling. Furthermore, even with valid instruments for number of years spent in an educational facility, this quantity may not perfectly reflect true education, a distinction pointed out at least as early as in Griliches (1977). In this paper, we follow Isacsson (1999) and use administrative data on educational attainment as an instrument for self-reported educational attainment in an attempt to mitigate the attenuation resulting from measurement error in schooling. As the data of this study present limited opportunity to examine the issue of mismeasured education, the twin methodology will be given the benefit of the doubt; the assumption of perfectly instrumented schooling will be maintained, and focus is instead directed towards the source of the alleged benefits from using twins data — the equal or virtually equal ability within twin pairs.

This paper is structured as follows. Section 2 sets out a simple theoretical framework which encompasses previous examinations and within which we propose two straightforward tests relying on less restrictive assumptions but which require richer data than has previously

[1] For critiques of the instrumental variable approach, see Bound, Jaeger, and Baker (1995) and Bound and Jaeger (1996).

[2] Isacsson (2004) distinguishes itself from the other papers in this list, as it develops an estimating framework to allow for non-classical measurement errors, and hence is able to provide a substantially more refined analysis than other specialist papers in this literature. Isacsson's estimates suggest that the classical measurement-error corrections are upwards biased by approximately 30%.

[3] For a review of the biological and developmental mechanisms that can give rise to differences between twins, see Martin et al. (1997).

[4] There is also a literature outside economics which reports associations between birthweight and educational attainment within twin pairs, see the review in Bound and Solon (1999).

been available. In Section 3 we describe such improved data, assembled by joining register data from a number of Swedish sources. Section 4 presents the results of various validity tests performed using this improved data, along with some robustness checks. Section 5 contains a discussion and Section 6 concludes.

2. Empirical framework

2.1. An augmented co-twin model

Consider the following simple model of wage determination, drawing on Card (1999):

$$y_{ij} = \alpha_y + \beta S_{ij} + \gamma A_{ij} + u_{ij}, \tag{1}$$

where y_{ij}, S_{ij} and A_{ij} are income in natural logarithms, years of schooling, and ability, respectively, for individual i of twin pair j, and where the ordering of the individuals in a twin pair is random. Returns to schooling, β, and the conditional return to ability, γ, are assumed to be equal across individuals. Let latent ability, A, be defined widely enough to allow S and u to be independent, and be measured in standard deviations about the population mean. Finally, α_y varies with a quadratic in the age of the individual, to capture experience and cohort-specific effects. Furthermore, assume the following causal model of schooling,

$$S_{ij} = \alpha_S + \delta A_{ij} + \epsilon_{ij}, \tag{2}$$

where ϵ is a summary measure of all determinants of schooling which are exogenous to the unobservables of the wage equation. Extend this exogeneity to apply across twins within a pair, so that $Corr(A_{ij},\epsilon_{kj}) = 0$ and $Corr(u_{ij},\epsilon_{kj}) = 0$, $\forall\ i, j$. Specify the sign of ability such that $\delta > 0$. Notice that this assumption is without loss of generality because A is not observed. Therefore, our approach does not make any assumption about the direction of the ability bias, as γ is free to take on either positive or negative values.

To capture cohort-specific effects, the intercept again varies with a quadratic function of age. Let the ability of a twin be statistically related to the ability of his co-twin in the following manner:

$$A_{1j} = \phi A_{2j} + \alpha_{1j}. \tag{3}$$

Here, ϕ is the correlation between the abilities of each twin and his co-twin, and α_{1j} is uncorrelated with A_{2j} by construction. Equivalently, ϕ is the share of variance in ability explained by a variance factor common to both twins. Furthermore, assume that differences in ability within pairs are independent of all other errors (u, ϵ, and τ (below)). The main identifying assumption of the literature on estimating the returns to schooling using variation within twin pairs, is that twins have identical latent abilities such that $A_{1j} = A_{2j}$. In the above framework, this translates to assuming $\phi = 1$, which in turn implies $Var(\alpha) = 0$ due to the random ordering of twins. Under $\phi = 1$, consistent estimates of β can be obtained by estimating the model in first-differences,

$$\Delta y_j = \beta_{FD}\Delta S_j + \gamma_{FD}\Delta A_j + \Delta u_j, \tag{4}$$

where $\Delta y_j \equiv y_{1j} - y_{2j}$ and similarly for the explanatory variables. Since ΔA_j is a zero vector under the standard twin assumption, the within-pair difference in income can simply be regressed on the within-pair difference in schooling,

$$\Delta y_j = \beta_{FD}^{-}\Delta S_j + \Delta u_j^{-}. \tag{5}$$

This is the basic idea behind all within-pair estimators in the literature. The aim of this study is to determine whether $\phi = 1$. For this purpose, consider IQ measured at around the age of 18, and specify its relationship with ability as follows:

$$T_{ij} = A_{ij} + \tau_{ij}, \tag{6}$$

where τ_{ij} is independent of A_{ij}. Let T_{ij} be measured in standard deviations about the population mean, and the unit of A_{ij} is defined implicitly. Finally, let y_1 refer to own income, as opposed to y_2 for co-twin's income, and similarly for S, A, T, u, ϵ, and τ, so that $(y_1)_{ij} = (y_2)_{kj}$; $\forall\ i \neq k$. When not specified, as above, y refers to own income, y_1.

2.2. Two tests of the basic twin assumption

Assume $Corr(u_1,\tau_1) = Corr(u_1,\tau_2) = 0$. Estimate the equation,

$$\Delta y_j = \beta\Delta S_j + \lambda_1\Delta T_j + \Delta u_j^{*}, \tag{7}$$

where the error term is,

$$\Delta u_j^{*} = -\lambda_1\Delta T_j + \gamma_{FD}\Delta A_j + \Delta u_j. \tag{8}$$

For our first test, we note that if $\phi = 1$, then $\Delta A_j = 0$ and $\Delta T_j = \Delta\tau_j$, and consequently $\lambda_1 = 0$. Furthermore, β and λ_1 are consistently estimated since $\lambda_1\Delta T_j = 0$ and $\gamma_{FD}\Delta A_j = 0$, and hence independent of ΔS_j and of ΔT_j. The distribution of $\hat{\lambda}_1$ is different under the null and the alternative hypothesis. It follows that a valid test statistic can be constructed from $\hat{\lambda}_1$ for the null hypothesis that $\phi = 1$. Measurement error in schooling can be dealt with using an alternative measure of schooling as an instrument, the approach championed in this literature by Ashenfelter and Krueger (1994), assuming, of course, that the exclusion restriction is satisfied.

For our second test, we note that the estimated return to education should change significantly when including IQ as a covariate in the fixed effects wage equation only if the equal ability assumption is invalid.[5] Denote the coefficient on schooling in the fixed effects regression without IQ included by β_1 and denote the coefficient on schooling in the regression with IQ included by β_2. A simple bootstrap procedure to test the hypothesis that the difference in estimated coefficients is not purely due to sampling variation is as follows. First, draw 10,000 pseudo-samples of twin pairs with replacement. For each bootstrap draw, estimate β_1 and β_2. An n-percent confidence interval for the quantity $\beta_1 - \beta_2$ can then be constructed by extracting the $\frac{n}{2}$ th and $\left(100-\frac{n}{2}\right)$ th percentile of the empirical distribution of $\beta_1 - \beta_2$ obtained from the bootstrap draws.

Note that under either of the two assumptions above, almost regardless of what exactly is measured by our IQ test, if its inclusion significantly affects the estimated coefficient on schooling, then the equal abilities assumption is violated. The co-twins literature assumes that apart from years of schooling, twins are identical. If there is a wage-relevant factor which is omitted and differs among twins, then the co-twins estimates are biased.

Furthermore, if differences in IQ is merely a determinant of schooling without having any direct impact on wage (only an indirect impact), then its inclusion should not affect the estimated return to schooling, under the assumptions of the existing co-twin literature, where schooling is perfectly instrumented.

3. Data

The dataset links information from the Swedish Twin Registry with administrative data from Statistics Sweden and Swedish enlistment records. The Swedish Twin Registry contains virtually all twins born in

[5] This implication does not however go the other way; i.e., it is not the case that a violation of the equal abilities assumption necessarily leads to coefficient changes when IQ is included as a covariate.

Sweden from 1926 and onwards, and is kept mainly for the purpose of performing epidemiological studies (see Lichtenstein et al. (2006) for a description of the Swedish twin registry). The survey data used in this paper was collected in 1998–2002 (the "SALT" survey) from twins born 1950–1958, and in 2005–2006 (the "STAGE" survey) from twins born in 1959–1975. Response rates to the two surveys were 74% and 60%, respectively. Only data on monozygotic twins (about one quarter of the sample) is used, where zygosity has been determined by the Swedish Twin Registry using a battery of questions relating to physical similarity. The validity of this method of determining zygosity has been repeatedly estimated to be 95–98% (Lichtenstein et al., 2002). The dataset is restricted to individuals born between 1950 and 1975.[6] The cohort studied is hence sufficiently old so that income is observed at a point in the lifecycle where research has shown that annual income is a good proxy for lifetime earnings (Böhlmark and Lindquist, 2006).

3.1. Education data

The data contains two measures of educational achievement. One is a self-reported measure from the survey data collected by the Swedish Twin Registry. The other is based on administrative data from 2005. The self-reported data consists, for the SALT cohort, of an indicator of highest attained qualification, and for the STAGE cohort, of total years of schooling at the different levels of the education system. For the SALT cohort, years of schooling are assigned based on the standard years of schooling associated with the degree in question. The administrative data contains highest degree attained. Years of schooling based on the survey data are used as the explanatory education variable, with degree dummies based on administrative sources used as instruments. In effect, this assumes classical measurement errors, an approximation which may not hold in the data, as demonstrated by Isacsson (2004) in a careful examination of earlier Swedish income data using only one measure for schooling. For our purposes, we nevertheless think this approximation is appropriate, as we wish to evaluate previous studies in this literature, none of which, except Isacsson (2004), have assumed non-classical measurement errors. We also note that the estimated impact of the non-classical measurement errors in Isacsson (2004) are derived under the assumption of equal abilities within twin pairs. In this sense Isacsson (2004) and the present study are complementary in that both studies relax one of the standard twin model assumption, but neither relaxes both.

3.2. Income data

Data on income consists of yearly taxable earnings in 2005 as reported by employers to the tax authorities. The income measure used in this paper ("sammanräknad förvärvsinkomst") is defined as the sum of income earned from wage labor, income from own business, pension income and unemployment compensation. Capital income is not included in the measure. In the main specification, only pairs where both twins in a pair had an income exceeding SEK 70,000 (exchange rate 2013; $\$1 \approx$ SEK 6.5) are included, in an attempt to capture only individuals working full-time so that income more or less corresponds to hourly earnings. The practice of either excluding data not corresponding to full-time work or using information on hourly wages is followed by practically all previous studies of the returns to schooling using twins back to at least Ashenfelter and Krueger (Ashenfelter and Krueger, 1994; Ashenfelter and Rouse, 1998; Behrman and Rosenzweig, 1999; Bonjour et al., 2003; Isacsson, 1999; Isacsson, 2004; Miller et al., 1995; Rouse, 1999; Zhang et al., 2007).[7]

It should also be noted that several papers in the literature (Behrman et al., 1994; Miller et al., 1995) use average earnings of the occupation in which an individual was employed as their measure of income. Our registry-based income measure, although by no means perfect, should nevertheless be a substantial improvement on these papers. Furthermore, several papers (Ashenfelter and Krueger, 1994; Ashenfelter and Rouse, 1998; Behrman and Rosenzweig, 1999; Bonjour et al., 2003; Zhang et al., 2007) use self-reported income, which will suffer from error if individuals lack perfect recall of their income levels, or if they do not define their income in exactly the same way. In fact, the only studies we are aware of that do not rely on either estimated earnings based on occupation or self-reported income are the two studies by Isacsson (1999, 2004) on Swedish data. In both those studies, a threshold of SEK 60,000 was imposed. We have increased this threshold slightly, as our income data is taken 15 years later than in Isacsson's (1999, 2004) two studies. Our threshold corresponds to an hourly wage of about \$5.65, and should be low enough to cover anyone with a full-time job. Finally, the exact level of the threshold is not important, as will be demonstrated in the section on robustness tests, where we re-examine our results using alternative thresholds of SEK 50,000 and 180,000. Lower levels than this are rarely used in empirical work, presumably because at such earnings or income levels, the numbers are not meaningful proxies of productivity.

3.3. IQ test score data

All Swedish men are required by law to participate in military conscription at or around the age of 18. Until 1999, exceptions were only granted to men with serious documented psychological or physical handicaps. The actual drafting procedure can span several days during which a number of tests are administered to the conscripts. These include assessments of medical status, physical stamina, muscular strength, eyesight, cognitive ability, and psychological aptitude. This paper takes a conservative approach and uses the data on cognitive ability, the most commonly used indicator for ability (see e.g. Hanushek and Woessman, 2008, for a discussion). As the normal school starting age in Sweden is seven, the average individual in the main sample would have taken the test about one year prior to finishing high school.

The IQ test used by the Swedish military is a fairly standard test of general intelligence (Spearman, 1904). An early version of the test was developed during World War II, and it has subsequently been revised on seven occasions (Carlstedt, 2000). Its basic structure has, however, remained unchanged during the study period considered in this paper. Recruits take four subtests: logical, verbal, spatial and technical. Carlstedt (2000) discusses the history of psychometric testing in the Swedish military and provides evidence on the psychometric properties of the test. The exact items used on the test are a military secret. Test scores are normalized by year using all observations in the dataset for which there are test scores, and the sum across test scores is then used as the raw IQ measure.[8] This raw measure is then normalized against all observations in the dataset, to allow for an approximation of population standard deviations to be used as the metric for IQ.

Using IQ test scores which were gathered not in a school environment, but under the considerably different conditions of military conscription, reduces the risk that the test scores pick up factors related to, i.e., a general affinity with school-like tests that yet do not translate into wage earning capacity. Using the terminology of the empirical framework outlined above, this renders it more plausible that $Corr(\epsilon_1,\tau_1)$ is zero. One of our robustness checks also involves taking the first principal component of the four cognitive tests to construct the IQ measure, thus only using variation common to the subtests when creating the variable (Spearman, 1904).

[6] 1950 is the first year for which we have data on IQ test scores.

[7] It can also be noted that due to the logarithmic transformation, some outliers in the full dataset are more than 10 standard deviations lower than the average.

[8] Assigning equal weights to each sub-test is in accordance with the standard practice of the Swedish Armed Forces.

3.4. Birth weight data

One of our robustness checks uses birth weight as an alternative measure of ability. The main source of these data is again the STAGE and SALT surveys which both contained the question "What was your birth weight?". For twins in the SALT cohort, data on birth weight has also been collected from delivery archives throughout Sweden (Lichtenstein et al., 2006). The birth records contain detailed information on the birth characteristics of each child, including birth weight. The archival data is preferable to the self-report data, so when both are available we use for archival data. The birth weight variable is standardized to mean zero and a standard deviation of one across the entire sample, to allow the results to be easily comparable with the results when using IQ as the proxy for ability.

3.5. Representativeness

The total sample size was determined as follows: Out of the 31,824 respondents to the STAGE and SALT surveys in our cohorts, 3522 were male monozygotic twins of which 2753 had data on education from both administrative and survey data. Of these, 2353 had non-missing income, and 2288 had an income above 70,000 SEK, the cut-off used to eliminate observations whose income unambiguously did not derive from full-time employment. Among these, 2129 individuals had valid IQ test scores from enlistment data. Finally, 1780 of these observations were from complete pairs of twins, i.e. where the co-twin was also in the sample. There were 1494 observations from complete male monozygotic pairs born between 1950 and 1975 where income data, the two measures of educational attainment and birth weight data was available.

Before turning to the main results, some comments on the representativeness of the sample are in order. In Table 1 the main sample is compared to the national average with regards to income, education, marital status and age. For IQ, the norm group is the approximately 12,000 twins born between 1950 and 1975 who responded to the SALT or STAGE survey and for whom there is IQ test score data. For all other variables, the comparison is made to the population data from Statistics Sweden. Income in the sample is about 20% higher than in the general population. Both education and age are slightly higher in the sample than in the national average, but these differences are small. Oversampling of twins with better than average education and income was also reported by Ashenfelter and Kruger (1994) and Ashenfelter and Rouse (1998).

It is also important to know how representative the dataset is of the datasets of twins used hitherto. Table 2 compares parameters from our dataset to parameters reported previously in the literature. The first two parameters concern similarity between twins. In our data, measured years of schooling correlate 0.73 between a twin and his co-twin, a figure in line with what has been reported in the literature. Furthermore, results on IQ test scores correlate 0.82, which again is a standard degree of similarity (Bouchard and McGue, 1981). The next two parameters concern the structure of the measurement errors in reported years of schooling. In our sample, the reliability ratio[9] is 0.88, which is very similar to those reported in previous twin studies. The reliability ratio of the within-pair differences is 0.65, which is closer to the lower than to the higher estimates reported in Ashenfelter and Krueger (1994) and Ashenfelter and Rouse (1998). The observed within-pair reliability ratio in the data is also close to that expected based on the cross-sectional reliability ratio and the twin correlation in schooling, as reported above. If all measurement errors are classical, the imputed within-pair reliability ratio would thus be 0.58.[10] Note also that the cross-sectional reliability ratio of 0.88 implies, under classical errors, a within-pair correlation in schooling of 0.82 (0.73/0.88) when correcting for measurement errors. As shown by Griliches (1979), co-twin estimators are less biased than cross-sectional estimators if and only if ϕ is greater than the similarity in schooling, i.e. in this dataset 0.82.

Table 1
Summary statistics and sample representativeness.

	Monozygotic twins	Population
Income (in SEK 1000)	360	298
S.D.	(228)	(…)
Schooling (in years)	12.9	12.2
S.D.	(2.6)	(…)
IQ	0.12	0.00
S.D.	(0.92)	(1.00)
Age (in years)	42.9	42.3
S.D.	(7.6)	(…)
1 if Married	0.51	0.45
S.D.	0.50	(…)
# Observations	1780	(…)

Note: The schooling data for the sample is based on self-reported education.
IQ is measured in standard deviations around a mean of zero, using all twins in the dataset as the standardization sample (12,366 observations in total).
All other population variables are constructed using data for the universe of Swedish men aged 30–55 years in 2005.

The final four parameters concern impacts on wages (in logarithms), and as such we would expect them to vary depending on institutional factors in the countries where they are measured. The first parameter, β_{IV}, is a simple cross-sectional estimate of the returns to schooling in our sample. The second parameter is the within family estimate of the return to schooling in the sample of MZ twins. In both cases, to try to adjust for measurement error, a full set of dummy variables on educational attainment based on the administrative data are used as instruments for self-reported educational attainment.

The estimated returns to schooling from cross-sectional data are slightly lower than those found in studies from US and UK, but slightly higher than those of Isacsson (1999) using Swedish twins. However, Isacsson's (1999) sample includes both men and women, whereas our estimates are for men only. Our data yields larger differences between within-pair estimates and cross-sectional estimates than what is commonly found in twin studies. Notice that the result from Isacsson (1999) was constructed using an imputed within-pair measurement error, and as such is not strictly comparable to the other figures which apply instrumental variables techniques to correct for measurement error.

The final two parameters in Table 2 concern the relationship of IQ with labor market outcomes. The standardized regression coefficient in a regression of log income on the IQ test score is 0.16, i.e. an increase in IQ of one standard deviation is associated with an increase in income of about 16% in our sample of monozygotic twins. Bowles and Gintis (2002), based on a meta-study of 24 studies on US data, report an average coefficient of 0.27. This discrepancy corresponds reasonably to differences in income dispersion between US and Sweden, as reported in Gottschalk and Smeeding (1997).

Finally, the correlation between self-reported schooling and measured IQ is 0.51, a figure roughly in line with the average of 0.55 reported by Neisser et al. (1996) in an authoritative report on the state of intelligence research. It should be noted that the latter figure is based on IQ test scores from early years, mainly primary school. The fact that the correlation with schooling is lower in our data suggests that simultaneity in test scores, whereby differences in schooling cause differences in IQ, is not a major concern.

4. Results

Before turning to the regression-based results, Fig. 1 shows the three most important bivariate relationships in our data. The upper panel is a scatter diagram of the intrapair difference in income (natural

[9] With classical measurement errors, the reliability ratio is the square root of the R^2 from a regression of measured years of schooling on its instruments. If there is only one instrument, this is equivalent to the correlation, as stated in the table.
[10] The imputed within-pair reliability ratio is $(r - Corr(S_1,S_2))/(1 - Corr(S_1,S_2))$, where r is the reliability ratio based on cross-sectional measures.

Table 2
Comparability with previous literature.

	Sample	Literature	Country/ies	Source
Co-twin similarity				
$Corr(S_1^1, S_2^1)$	0.73	0.66	US	Ashenfelter and Kruger (1994)
		0.75	US	Ashenfelter and Rouse (1998)
		0.70	Australia	Miller et al. (1995)
$Corr(T_1, T_2)$	0.82	0.86	Various	Bouchard and McGue (1981)
Education instruments				
Cross-sectional reliability ratio	0.88	≈0.90	US	Ashenfelter and Kruger (1994)
		0.92	US	Ashenfelter and Rouse (1998)
		0.88	Australia	Miller et al. (1995)
		0.88	Sweden	Isacsson (1999)
Within-family reliability ratio	0.65	0.57–0.83	US	Ashenfelter and Kruger (1994)
		0.62–0.76	US	Rouse (1999)
Labor market				
β_{IV}	7.2%	≈ 8%	US, UK	Card (1999), Bonjour et al. (2003)
		6.4%	Australia	Miller et al. (1995)
		5.2%	Sweden	Isacsson (1999)
$\beta_{FE,IV}$	3.4%	≈ 7%	US, UK	Card (1999), Bonjour et al. (2003)
		4.5%	Australia	Miller et al. (1995)
		4.2%	Sweden	Isacsson (1999)
$\delta y/\delta T$	0.16	0.27	US	Bowles and Gintis (2002)
$Corr(S^1, T)$	0.51	0.55	US	Neisser et al. (1996)

Note: S^1 is self reported schooling, T is measured IQ, y is log income. Subscripts refer to a twin's order in a pair.
The sample "correlation" of schooling and instrument for schooling was derived as the square root of the R^2 when regressing self-reported years of schooling on the set of administrative schooling dummies used as instruments. β_{IV} is the regression coefficient from the cross-sectional regression of log income on schooling (S^1), using a set of dummies on educational attainment categories from administrative records as instruments. $\beta_{FE,IV}$ is the within-family estimate of the return to schooling, using the within pair difference in the set of dummy variables as instruments. $\delta y/\delta T$ is the standardized regression coefficient in the regression of log income on measured IQ.

logarithm) against the intrapair difference in schooling. It is clear from the figure that a large number of identical twins do indeed have identical levels of educational attainment, and that within-pair variation in educational attainment is associated with within-pair variation in earnings. The middle panel plots the intrapair difference in IQ against the intrapair difference in schooling and shows that the relationship is positive. The bottom panel shows that there is also a positive relationship between IQ and income within pairs, i.e. when examining the first-order relationship without controlling for variation in schooling. In examining these figures, it is useful to recall that the signal-to-noise ratio is lower within pairs than it is in the cross-section. Assuming classical measurement error and a cross-sectional reliability ratio for IQ of 0.9 implies a measurement-error corrected within-pair correlation of educational attainment and IQ of 0.30. [11] This number in and of itself casts doubts on the co-twin methodology.

4.1. Main results

Table 3 reports our main results from specifications with and without IQ test scores included. All regressions have family fixed effects, so the only source of variation is the within-family differences. Standard errors are clustered at the family level. The first two columns report the results from a model estimated by OLS. Columns 3 and 4 report the results from a model in which the self-reported schooling variable is instrumented for using the administrative dummies. To maintain comparability with earlier work, we also report results from a specification in which a continuous schooling variable is used as the instrument (see columns 5–6). To construct the continuous measure, we use Isacsson's (2004) estimates of the average educational attainment associated with each of the administrative categories that define the dummy variables. [12]

The results clearly show that within-pair differences in IQ have a direct relationship with income differences, and that this relationship is statistically significant and strong. The results are quite similar across the three specifications, so we focus on the specification where the administrative dummies are used as instruments. The magnitude of the coefficient implies that a twin with an IQ one population standard deviation higher than his co-twin, has an income which is on average 7.4% higher than his co-twin, despite controlling for schooling. The coefficient on schooling drops from 3.4% to 2.8%, or by about 15%.[13] Under the assumptions underpinning this test ($Corr(u_1,\tau_1) = Corr(u_1,\tau_2) = 0$), the hypothesis that $\phi = 1$ is hence rejected. Furthermore, using the previously described bootstrapping procedure, the null hypothesis that the schooling coefficients are the same in the specifications with and without IQ included can be rejected at the one percent level (p-value <0.01). The results from the bootstrapping procedure are reported in Table 3.

5. Robustness

There are a number of legitimate concerns which may be raised with regards to the findings in Table 3. In this section, we conduct six separate robustness checks of our main findings. The results from these analyses are summarized in Table 4.

5.1. Misclassification of twins

Some of the twins in the sample may have been misclassified as monozygotic twins despite being in fact dizygotic twins. If ability differences are for some reason relatively less familial (i.e., compared to the family share of variance of the exogenous determinant of schooling) in dizygotic twins, this will cause the above findings to be overstated. To examine this issue, the 5% of pairs which were the most dissimilar with respect to IQ were dropped and the main equations were re-estimated. This is a conservative test in that no more than 2–5% of monozygotic twins are normally misclassified as dizygotic using the type of classification algorithm employed by the Swedish Twin Registry (Lichtenstein *et al.* (2002)). In column 1 of Table 4, labeled "*Exclude 5%*",

[11] The imputed within-pair correlation is derived as $\frac{Corr(\Delta S^1, \Delta T)}{\sqrt{\rho_{\Delta S^1}}\sqrt{\rho_{\Delta T}}}$, where $\rho_{\Delta S^1}$ and $\rho_{\Delta T}$ are the reliability ratios of the two respective first-differenced variables, as derived in footnote 8.

[12] Isacsson (2004) examined a representative sample with high quality data on years of schooling and regressed this on the same type of administrative data that are used in this paper.

[13] It should be noted that since T is an imperfect measure of ability, the estimated returns to schooling are biased and inconsistent when $\phi \neq 1$, i.e. when the equal ability assumption is violated.

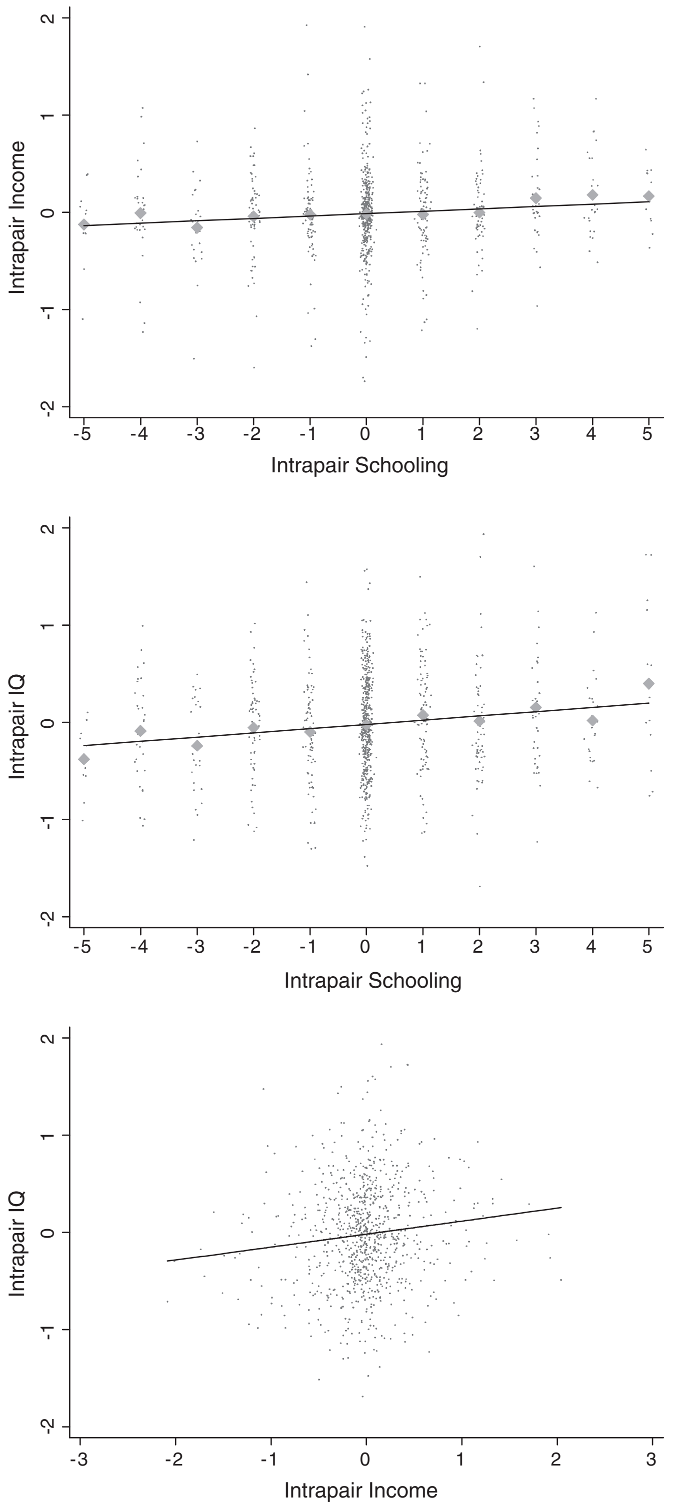

Fig. 1. Bivariate relationships. Upper Panel: Intrapair difference in ln income plotted against intrapair difference in schooling. Middle Panel: Intrapair difference in IQ plotted against intrapair difference in schooling. Lower Panel: Intrapair difference in IQ plotted against income.

we report the results from this robustness analysis. We find that the estimates in this subsample are very similar to those in our main specification; the coefficient of IQ on income is a little higher than in the main specification (0.100 versus 0.074), and the estimated return to schooling falls from 0.030 to 0.025 when IQ is included as a control. This fall is statistically significant (p-value <0.01).

Table 3
Results of the two tests of the equal ability assumption — main case.

	(1)	(2)	(3)	(4)	(5)	(6)
	FE	FE	FE/IV	FE/IV	FE/IV	FE
Dependent variable	Income	Income	Income	Income	Income	Income
Schooling	0.024***	0.021**	0.034***	0.028**	0.038***	0.033**
S. E.	(0.008)	(0.008)	(0.013)	(0.013)	(0.014)	(0.014)
IQ		0.078***		0.074***		0.072***
S. E.		(0.026)		(0.026)		(0.026)
Schooling instrument	None	None	Discrete	Discrete	Cont.	Cont.
Family fixed effects?	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.011	0.020	0.009	0.019	0.011	0.017
# Observations	1780	1780	1780	1780	1780	1780
Groups	890	890	890	890	890	890

Note: Standard error within parentheses, clustered at the family level. Three stars (***) denote statistical significance at the 1% level, two stars (**) denote significance at the 5% level and one star (*) denotes statistical significance at the 10% level. "Discrete": administrative dummies for highest degree attained are used as instruments for years of schooling. "Cont.": a continuous measure of years of schooling imputed from administrative dummies is used as the instrument for self-reported years of schooling.

5.2. High school restriction

The IQ measure is taken at about the age of 18, after individuals have completed compulsory schooling but before they enter college. The fact that the IQ tests are taken at a relatively early age renders it less likely that the differences in test scores are endogenous to differences in educational attainment. Yet, there is evidence suggesting that differences in education can drive differences in test scores (Cascio and Lewis, 2006). The argument that test scores are endogenous to differences in acquired human capital is particularly compelling for twin pairs where at least one twin has less than 12 years of schooling and hence either failed to complete high school or only completed a two-year high school curriculum. As a crude robustness check, we therefore restrict the sample to individuals whose education was still ongoing when they took the test, and rerun the analyses.[14] Column 2 of Table 4 ("*High School*") shows the results omitting twin pairs where at least one sibling failed to complete three years of high school. Again, the coefficient on schooling falls significantly (from 4.1 to 3.4%, p-value < 0.01) and IQ is a significant predictor of income even conditioning on schooling.

5.3. Alternative IQ measure

To examine the sensitivity of our findings to variations in the construction of the aggregate test score, a so called factor " g ", *i.e.* the first principal component, was calculated from the four subtests of the IQ test. This measure was standardized by year against all twins for whom there was data on IQ, and used as an alternative measure of IQ. The results for the alternative measure of IQ are reported in column 3 of Table 4, labeled "*Alt IQ*". These results are also highly similar to those in the main specification.

5.4. Instruments interchanged

As a further robustness check, the roles of instrument and regressor were reversed for the two sources of schooling data. As the administrative data, which were used as instruments in the main analysis, consist of dummy variables for highest degree attained, they were converted into years of schooling using Isacsson's imputation model (2004). We then instrumented for the imputed values using the self-reported years of schooling measure. We report the results from regression

[14] In principle, it is also possible that that differences in IQ test scores are caused by differences in program choices in high school. Unfortunately, we do not have any data that would allow us to examine this possibility.

Table 4
Robustness checks of the fixed effects regressions.

	(1) Exclude 5%	(2) High School	(3) Alt IQ	(4) Regr ⇆ Instr	(5) Threshold 50 k	(6) Threshold 180 k	(7) Trimming	(8) Birth weight
Panel A (no ability controls)								
Schooling	0.030**	0.041*	0.034**	0.035***	0.034**	0.026**	0.041**	0.043***
S. E.	(0.015)	(0.022)	(0.013)	(0.012)	(0.013)	(0.011)	(0.019)	(0.015)
R^2	0.002	0.007	0.009	0.007	0.011	0.005	0.004	0.012
Panel B (ability controls)								
Schooling	0.025*	0.034	0.028**	0.031**	0.029**	0.023**	0.034*	0.041***
S. E.	(0.014)	(0.022)	(0.013)	(0.012)	(0.014)	(0.012)	(0.019)	(0.015)
IQ	0.100***	0.129***	0.076***	0.071*	0.071***	0.059***	0.074***	–
S. E.	(0.034)	(0.042)	(0.026)	(0.027)	(0.027)	(0.023)	(0.026)	–
BW	–	–	–	–	–	–	–	0.036*
S. E.	–	–	–	–	–	–	–	(0.021)
R^2	0.014	0.024	0.019	0.017	0.019	0.015	0.014	0.017
# Observations	1692	906	1780	1780	1790	1582	1708	1494
p-Value	<1%	<1%	<1%	<2%	<2%	< 3%	<1%	<22%

Note: This table summarizes the results of the robustness checks. Standard errors are clustered at the family level. Three stars (***) denote statistical significance at the 1% level, two stars (**) denote significance at the 5% level and one star (*) denotes significance at the 10% level. All specifications include family fixed effects and administrative dummies for highest degree attained are used as instruments for years of schooling. Birth weight and IQ test scores are normalized to have a mean of zero and a standard deviation of one. "Exclude 5%": omits 5% of twins most dissimilar on IQ. "Exclude 5%": only keeps pairs in which both twins completed high school. "Alt IQ": IQ is defined as the principal component of the four cognitive subtests. "Regr ⇆ Instr": schooling instruments and regressors interchanged. "Threshold 50 k": earnings threshold set at 50,000 "Threshold 180 k": earnings threshold set at 180,000. "Trimming": omits pairs in which the difference in schooling exceeds four years. "Birth weight": birth weight used as ability measure. The p-value is for the test of the hypothesis that the return to schooling does not change when the ability measure is included as a control.

models with the instruments interchanged in column 4 of Table 4 ("*Regr ⇆ Instr*"). Again, the results are substantively identical to those in the baseline specification.

5.5. Full-time threshold

The sensitivity of the main results to variations in the threshold on yearly earnings was examined, by applying alternative thresholds of 50,000 and 180,000 Swedish krona (about \$6700 and \$24,000, respectively). Regarding the lower threshold, it should be noted that it corresponds to a full-time hourly wage of about \$3.4, *i.e.* implausibly low in the context of Sweden. Furthermore, because of the logarithmic conversion of wages, the 24 observations below the lower threshold are between 4 and 10 standard deviations away from the mean (in a sample of around 2000). The lower threshold is indeed very low for the purposes of approximating a full-time proxy and at lower thresholds than this, the income is arguably not a meaningful measure of productivity. The results are shown in columns 5 and 6 of Table 4 ("*Treshold 50 k*" and "*Threshold 180 k*"). For both the high and the low threshold, the estimated return to schooling falls by approximately ten percent when IQ test scores are included as controls and the fall is significant at the five percent level in both cases.

5.6. Trimming

We next examine if the results are sensitive to outliers. We defined an outlier pair as any pair in which the two siblings differ by more than 4 years of educational attainment. The results are shown in column 7 ("*Trimming*") and are very similar to those in our main specification; the estimated return to schooling falls from 4.1 to 3.4% (p-value < 0.01).

5.7. Birth weight

As a final robustness check we replace the IQ test scores by birth weight. The birth weight variable is of course immune to reverse causality concerns, but probably more weakly associated with ability. In the model with birth weight used as the ability measure, the difference in coefficients is significant at the ten percent level, though the bootstrapped difference in coefficients is not significant. These results are shown in column 8 ("*Birth Weight*").

5.8. Summary

In all cases, the IQ test score variable is statistically and economically significant in both the wage equation and the schooling equation. The estimated return to schooling also declines when IQ is included as a covariate. Finally, for all alternative samples, the null hypothesis that the schooling coefficients are the same in the specifications with and without IQ included can be rejected at the five percent level. The difference is significant at the five percent level in all cases. The results for birthweight are qualitatively similar but do not allow us to statistically reject the null hypothesis that the estimated return to schooling is the same with and without birth weight included as a control.

6. Discussion

In the previous literature, the main way of evaluating the equal abilities assumption has been to compare the across-pair correlation between schooling and a proxy for ability with the corresponding within-pair correlation. For example, Ashenfelter and Rouse (1998) provide evidence that in their sample, the correlation between on the one hand schooling and on the other hand the following proxies for ability – marital status, years of job tenure, father's education, mother's education, and spouse's education – were all significant at the one percent level in the cross-section, but not significant even at the ten percent level when comparing within pairs. A potential problem with this approach is that it is well known that measurement error in both schooling and the proxies would tend to be exacerbated when comparing within-pairs, so that within-pair correlations are not easily comparable with correlations in the cross-section. In addition, the analysis implicitly relies on the assumption that the proxy is uncorrelated with the non-ability components of schooling. Failure of this assumption will bias the test – quite possibly in the direction of failing to reject the null – and renders the interpretation complicated.[15]

Other authors who have performed the same type of analyses with qualitatively the same results include Bonjour et al. (2003) (birth weight, marital status, part-time work, partner's characteristics, adult

[15] In the case of the proxies proposed by Ashenfelter and Rouse (1998), this assumption is almost certainly violated, as the proxies are either taken at adult years and are therefore likely to be causally affected by schooling (marital status and years of job tenure), or are measures of father's, mother's, or spouse's education, measures which all would appear to be correlated with the non-ability component of schooling.

reading scores, and smoking in adolescence) and Miller et al. (2005) (birth weight). It is noteworthy however, that more recent studies which are based on significantly larger samples find significant within-pair correlations between birth weight and schooling (Behrman and Rosenzweig, 2004; Black et al., 2007; Royer, 2009).[16]

Isacsson (1999) performs a similar comparison by comparing cross-sectional and within-pairs regressions of schooling on psychological test measures (short forms of the Eysenck scale) and physiological variables (height and weight) taken between 14 and 20 years of age. Only the physiological variables have a significant effect in the cross-section, and consistent with the evidence cited above, these physiological variables have substantially smaller and statistically insignificant effects within pairs. Height and weight are however likely to be relatively poor proxies for ability (i.e. relatively high error as compared to a perfect ability measure), and as such they will be more sensitive to the exacerbation in measurement errors which follows from differencing within pairs.

The only directly comparable finding that we are aware of is in Griliches (1979), who reports a regression coefficient of 0.13 for the within-pair effect of one standard deviation in IQ on years of schooling, based on a small sample of just 76 pairs of male monozygotic twins from Project Talent. In his data, the within-pair correlation between IQ and schooling is merely 0.05, which is statistically indistinguishable from our point estimate of 0.15. With a sample size of 76, the statistical power to statistically reject at the five percent level the null hypothesis of a zero correlation when the true correlation is 0.15 is 25%. Hence, our estimates are in no way contradictory to those of Griliches (1979) and more generally, our conjecture is that the failure to reject the null hypothesis in the previous literature has been largely driven by low statistical power. The low power is likely explained by attenuation due to measurement error when computing within pair correlations and the choice of proxies which are only weakly related to ability.

The main finding in the previous sections is that the assumption of equal ability within pairs of monozygotic twins is violated in our sample. Within-pair variation in IQ test scores predicts within-pair variation in schooling, and including within-pair variation in IQ in the fixed effects regressions lowers the estimated return to schooling by approximately fifteen percent. This evidence against the equal ability assumption relies on the assumption that the discrepancy between the proxy and the true ability (τ) is uncorrelated with the unobservable portion of income which is uncorrelated with schooling and ability (u). This is likely to be a much weaker assumption than what has been implicit in previous studies, namely that (τ) is uncorrelated with the non-ability component of schooling (ϵ).

The results of this paper are robust across different alternative samples and likely understate the extent of the bias, for two reasons. First, inclusion of IQ does nothing to remedy the likely imperfections in the instruments used for true schooling, so that the various returns to schooling estimated within pairs are potentially all subject to exacerbated measurement error. Second, even if the schooling instruments are valid, the estimated decline in the schooling coefficient does not take into account concerns about the validity of IQ tests as measures of actual ability. This is of course all the more pertinent as differencing of the IQ test scores will exacerbate the problem of errors in variables under plausible assumptions about the measurement error process (Griliches, 1979).

Though we have reported evidence suggesting that the co-twin approach to estimating the returns to schooling produces biased estimates, it does not necessarily follow that the entire enterprise should be abandoned. For example, in their otherwise quite critical assessment of the co-twin method, Bound and Solon (1999, pp. 176–179) suggest that although we do not know whether ability is more familial than is schooling, within-pair estimates can still be used as an upper bound on the returns to schooling, under the assumption that ability bias is positive as is commonly thought (Bound and Solon (1999)).[17]

Given that within-pair IV estimates are generally lower than the cross-sectional OLS estimates, co-twin estimates then contain information allowing us to tighten the bounds on the possible values of the returns to schooling. However, the central premises of this type of bounds argument, that ability bias can be taken to be positive a priori and that the suitability of an identification method therefore can be determined on the basis of the results it provides – if lower than OLS, then accept as an improvement – can be criticized from a methodological perspective. Furthermore, as Bound and Solon (1999) note, such reasoning naturally rests on the assumption that the instruments for schooling as a measure of human capital are valid. Such an assumption is far from innocuous, and the potential reduction in bias must be weighed carefully against the plausibility of this assumption. Absent a fuller understanding of the exact shape of both the relationship between years of education and log incomes and of the properties of the error terms, our intuition is that the costs from using within-pairs variation – regardless of the availability of controls for ability – will often exceed the gain. At the very least, the results reported here suggest that it was premature of Greene to claim that the co-twin approach "ameliorates" (Greene, 2003, p. 381) the problems with standard cross-sectional regressions or of Böckerman and Vainiomaki (2013, p. 86) to assert that "twin data removes unobserved cognitive and non-cognitive ability differences".

7. Conclusion

Monozygotic twins' schooling decisions have been used in a number of prominent papers to estimate the returns to schooling. The key identifying assumption in these studies is that within-pair variation in schooling is explained by factors which are unrelated to wage earning ability. Using a unique dataset of 890 pairs of male monozygotic twins' schooling, income and adolescent IQ test scores, this paper finds strong evidence against the *equal ability assumption*. Within-pair differences in IQ test scores, obtained around the age of eighteen, are found to be a significant predictor of income even when controlling for schooling differences. Introducing within-pair IQ differences in a standard within-pair wage equation reduces the estimated returns to schooling by about 15%. The results are similar in magnitude when using birth weight as the proxy for ability.

References

Amin, Vikesh, 2011. Returns to education: evidence from UK twins: comment. Am. Econ. Rev. 101 (4), 1629–1635.

Angrist, Joshua D., Krueger, Alan B., 1991. Does compulsory school attendance affect schooling and earnings? Q. J. Econ. 106 (4), 979–1014.

Ashenfelter, Orley, Krueger, Alan B., 1994. Estimating the returns to schooling using a new sample of twins. Am. Econ. Rev. 84 (5), 1157–1173.

Ashenfelter, Orley, Rouse, Cecilia, 1998. Income, schooling and ability: evidence from a new sample of identical twins. Q. J. Econ. 113 (1), 253–284.

Becker, Gary, S., 1964. Human capital: a theoretical and empirical analysis, with special reference to education. Columbia University Press for the National Bureau of Economic Research, New York.

Behrman, Jere R., Rosenzweig, Mark, 1999. 'Ability' biases in schooling returns and twins: a test and new estimates. Econ. Educ. Rev. 18 (2), 159–167.

Behrman, Jere R., Rosenzweig, Mark, 2004. Returns to birth weight. Rev. Econ. Stat. 86 (2), 586–601.

Behrman, Jere R., Rosenzweig, Mark, Taubman, Paul, 1994. Endowments and the allocation of schooling in the family and in the marriage market: the twins experiment. J. Polit. Econ. 102 (6), 1131–1174.

Ben-Porath, Yoram, 1967. The production of human capital and the life cycle of earnings. J. Polit. Econ. 75 (4), 352–365.

Black, Sandra E., Devereux, Paul J., Salvanes, Kjell G., 2007. From the cradle to the labor market? The effect of birth weight on adult outcomes. Q. J. Econ. 122 (1), 409–439.

Böckerman, Petri, Vainiomaki, Jari, 2013. Stature and life-time labour market outcomes: accounting for unobserved differences. Labour Econ. 24 (1), 86–96.

[16] Bonjour et al. (2003) do not infer that the equal ability assumption holds, but merely that the ability bias should be reduced by using within-pair variation.

[17] We emphasize that Bound and Solon (1999) do not strictly advocate the co-twin approach, but merely point out this logical implication.

Böhlmark, Anders, Lindquist, Matthew J., 2006. Life-cycle variations in the association between current and lifetime income: replication and extension for Sweden. J. Labor Econ. 24 (4), 879–900.
Bonjour, Dorothe, Cherkas, Lynn F., Haskel, Jonathan E., Hawkes, Denise D., Spector, Tim D., 2003. Returns to education: evidence from UK twins. Am. Econ. Rev. 93 (5), 1799–1812.
Bouchard Jr., Thomas J., McGue, Matt, 1981. Familial studies of intelligence: a review. Science 212 (4498), 1055–1059.
Bound, John, Jaeger, David A., 1996. On the validity of season of birth as an instrument in wage equations: a comment on Angrist and Krueger's 'Does Compulsory School Attendance Affect Schooling and Earnings? NBER Working Paper No. 5835.
Bound, John, Solon, Gary, 1999. Double trouble: on the value of twins-based estimation of the return to schooling. Econ. Educ. Rev. 18 (2), 169–182.
Bound, John, Jaeger, David A., Baker, Regina M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. J. Am. Statist. Assoc. 90 (430), 443–450.
Bowles, Samuel, Gintis, Herbert, 2002. The inheritance of inequality. J. Econ. Perspect. 16 (3), 3–30.
Card, David, 1999. the causal effect of education on earnings. In: Ashenfelter, Orley, Card, David (Eds.), *Handbook of labor economics*, vol. 3A. Elsevier, Amsterdam, NL.
Card, David, 2001. Estimating the return to schooling: progress on some persistent econometric problems. Econometrica 69 (5), 1127–1160.
Carlstedt, Berit, 2000. Cognitive Abilities: Aspects of Structure, Process and Measurement. University of Gothenburg, Göteborg, Sweden (PhD dissertation).
Cascio, Elizabeth U., Lewis, Ethan G., 2006. Schooling and the armed forces qualifying test. J. Hum. Resour. 41 (2), 294–318.
Gottschalk, Peter, Smeeding, Timothy M., 1997. Cross-national comparisons of earnings and income inequality. J. Econ. Lit. 35 (2), 633–687.
Greene, William H., 2003. Econometric Analysis: Fifth Edition. Prentice-Hall Inc, Upper Saddle River, New Jersey.
Griliches, Zvi, 1977. Estimating the returns to schooling: some econometric problems. Econometrica 45 (1), 1–22.
Griliches, Zvi, 1979. Sibling models and data in economics: beginnings of a survey. J. Polit. Econ. 87 (5), S37–S65.
Hanushek, Eric A., Woessman, Ludger, 2008. The role of cognitive skills in economics development. J. Econ. Lit. 46 (3), 607–688.
Isacsson, Gunnar, 1999. Estimates of the return to schooling in sweden from a large sample of twins. Labour Econ. 6 (4), 471–489.
Isacsson, Gunnar, 2004. Estimating the economic return to educational levels using data on twins. J. Appl. Econ. 19 (1), 99–119.
Katz, Lawrence, Autor, David, 1999. Changes in the wage structure and earnings inequality. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 3A. Elsevier, Amsterdam, NL.
Lichtenstein, Paul, de Faire, Ulf, Floderus, Birgitta, Svartengren, Magnus, Svedberg, Pia, Pedersen, Nancy L., 2002. The Swedish twin registry: a unique resource for clinical, epidemiological and genetic studies. J. Intern. Med. 252 (3), 184–205.
Lichtenstein, Paul, Sullivan, Patrick F., Cnattingius, Sven, Gatz, Margaret, Johansson, Sofie, Carlström, Eva, Björk, Camilla, Svartengren, Magnus, Volk, Alicja, Klareskog, Lars, de Faire, Ulf, Schalling, Martin, Palmgren, Juni, Pedersen, Nancy L., 2006. The Swedish twin registry in the third millennium: an update. Twin Res. Hum. Genet. 9 (6), 875–882.
Martin, Nicholas, Boomsma, Dorret, Machin, Geoffrey, 1997. A twin-pronged attack complex traits. Nat. Genet. 17, 387–392.
Miller, Paul W., Mulvey, Charles, Martin, Nicholas, 1995. What do twins studies reveal about the economic returns to education? a comparison of Australian and U.S. findings. Am. Econ. Rev. 85 (3), 586–599.
Miller, Paul W., Mulvey, Charles, Martin, Nicholas, 2005. Birth weight and schooling and earnings: estimates from a sample of twins. Econ. Lett. 86 (3), 387–392.
Miller, Paul W., Mulvey, Charles, Martin, Nicholas, 2006. The return to schooling: estimates from a sample of young Australian twins. Labour Econ. 13 (5), 571–587.
Neisser, Ulric, Boodoo, Gwyneth, Bouchard Jr., Thomas J., Boykin, Wade, A., Brody, Nathan, Ceci, Stephen J., Halpern, Diane F., Loehlin, John C., Perloff, Robert, Sternberg, Robert J., Urbina, Susana, 1996. Intelligence: knowns and unknowns. Am. Psychol. 51 (2), 77–101.
Neumark, David, 1999. Biases in Twin Estimates of the Return to Schooling. Econ. Educ. Rev. 18 (2), 143–148.
Rouse, Cecilia, 1999. Further estimates of the economic return to schooling from a new sample of twins. Econ. Educ. Rev. 18 (2), 149–157.
Royer, Heather, 2009. Separated at girth: US twin estimates of the effects of birth weight. Am. Econ. J. Appl. Econ. 1 (1), 49–85.
Spearman, Charles, 1904. General intelligence, objectively determined and measured. Am. J. Psychol. 15, 201–292.
Taubman, Paul, 1976. Earnings, education, genetics, and environment. J. Hum. Resour. 11 (4), 447–461.
Zhang, Junsen, Liu, Pak-Wai, Yung, Linda, 2007. The cultural revolution and returns to schooling in China: estimates based on twins. J. Dev. Econ. 84 (2), 631–639.